

Author: Dmitri Zaykin

Investigator

Biostatistics Branch

National Institute of Environmental Health Sciences

Title: Determining the number of SNPs for following up large scale association studies.

Abstract: In recent years, genome-wide association studies (GWAS) have uncovered a large number of susceptibility variants. Nevertheless, GWAS findings provide only tentative evidence of association, and replication studies are required to establish their validity. Due to this uncertainty, researchers often focus on top ranking SNPs, instead of considering strict significance thresholds to guide replication efforts. The number of SNPs for replication is often determined ad hoc. We show how the rank-based approach can be used for sample size allocation in GWAS as well as for deciding on a number of SNPs for replication. The basis of this approach is the "ranking probability": chances that at least Y true associations will rank among top Z SNPs, when SNPs are sorted by P-value. We accommodate linkage disequilibrium and relate ranking probabilities to the proportion of false discoveries among top SNPs. A study-specific proportion can be estimated from P-values, and its expected value can be predicted for study design applications.